Group 1
Anders Enqvist
Firooz Shahriari Mehr
Ozan Alp Topal
Yasaman Khorsandmanesh

Homework 1 Solutions

HW1.1: prove all the statements

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{\mu}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2 \qquad (4)$$

1 - (4) is equivalent to a minimum positive curvature $\nabla^2 f(\boldsymbol{x}) \geq \mu \boldsymbol{I}_d, \forall \boldsymbol{x} \in \mathcal{X}$

2 - (4) is equivalent to $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \mu \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$

3 - (4) implies

(a) **Polyak-Łojasiewicz (PL) Inequality**: $f(\boldsymbol{x}) - f^\star \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2, \forall \boldsymbol{x}$

(b) $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2 \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2$

(c) $(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2$

(d) $f(\boldsymbol{x}) + r(\boldsymbol{x})$ is strongly convex for any convex f and strongly convex r

1- let $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2 \Rightarrow \nabla\phi(x) = \nabla f(x) - \mu x$

as $f(x)$ is μ-strongly convex $\Rightarrow$ $\quad + \frac{\mu}{2}\|x_1\|_2^2 - \frac{\mu}{2}\|x_1\|_2^2$

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T(x_2 - x_1) + \frac{\mu}{2}\left(\|x_2\|_2^2 - 2x_1^T x_2 + \|x_1\|_2^2\right)$$

$$\Rightarrow f(x_2) - \frac{\mu}{2}\|x_2\|_2^2 \geq f(x_1) - \frac{\mu}{2}\|x_1\|_2^2 + (\nabla f(x_1) - \mu x_1)^T(x_2 - x_1)$$

$\Longleftrightarrow$ which implies convexity of $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$

$$\Longleftrightarrow \nabla f(x) - \mu x \succeq 0 \Longleftrightarrow \nabla^2 f(x) - \mu I \succeq 0$$

$$\Longleftrightarrow \nabla^2 f(x) \succeq \mu I \qquad \square$$

2- From equation 4, we can write

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|_2^2$$

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$

$$\overset{+}{\Longrightarrow} \quad 0 \geq (\nabla f(x_1) - \nabla f(x_2))^T (x_2 - x_1) + \mu \|x_2 - x_1\|_2^2$$

$$\Longrightarrow \quad (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \mu \|x_2 - x_1\|_2^2 \quad \square$$

3-a) by taking minimization respect to x_2 on both sides of (4). Then we have

$$\min_{x_2} \{f(x_2)\} = f(x^*) = f^* \quad (\text{left hand side})$$

$$\frac{\partial}{\partial x_2} (\text{right hand side}) = \nabla f(x_1) + \mu (x_2 - x_1) = 0 \Longrightarrow$$

$$x_2 = x_1 - \frac{1}{\mu} \nabla f(x_1) \xrightarrow[\text{side}]{\text{right hand}} f(x_1) - \frac{1}{\mu} \|\nabla f(x_1)\|^2 + \frac{1}{2\mu} \|\nabla f(x_1)\|_2^2$$

$$\xrightarrow{\text{is valid for } \forall x} \quad f^* \geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|_2^2 \quad \square$$

3-b) by using $|ab| \leq |a||b|$ and what we proved in 2

$$\|\nabla f(x_2) - \nabla f(x_1)\|_2 \|x_2 - x_1\| \geq \left|(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1)\right| \geq \mu \|x_2 - x_1\|_2^2$$

rearranging

$$\Rightarrow \|x_2 - x_1\|_2 \leq \frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2 \qquad \square$$

3-c) let's define $\phi(x_3) = f(x_3) - \nabla f(x_1)^T x_3$, where it is μ-strongly convex since

$$(\nabla \phi(y) - \nabla \phi(z))^T (y - z) = (\nabla f(y) - \nabla f(z))^T (y - z) \geq \mu \|y - z\|_2^2$$

by applying (3-a) result to

$\Longrightarrow$

strongly convex property $\phi(x_3)$ with $x_3^* = x_1$

$$(f(x_2) - \nabla f(x_1)^T x_2) - (f(x_1) - \nabla f(x_1)^T x_1) = \phi(x_2) - \phi(x_1)$$

$$\leq \frac{1}{2\mu} \|\nabla \phi(x_2)\|^2 = \frac{1}{2\mu} \|\nabla f(x_2) - \nabla f(x_1)\|^2 \,,$$

$$\Rightarrow \begin{cases} f(x_2) \leq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2^2 \\ f(x_1) \leq f(x_2) + \nabla f(x_2)^T (x_1 - x_2) + \frac{1}{2\mu} \|\nabla f(x_1) - \nabla f(x_2)\|_2^2 \end{cases}$$

$$\overset{+}{\Longrightarrow} (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \leq \frac{1}{\mu} \|\nabla f(x_2) - \nabla f(x_1)\|_2^2 \qquad \square$$

3-d)

$f(x)$ is convex $\Rightarrow f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1)$

$r(x)$ is μ-strongly convex $\Rightarrow r(x_2) \geq r(x_1) + \nabla r(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$

$\xRightarrow{+} f(x_2) + r(x_2) \geq f(x_1) + r(x_1) + \left(\nabla f(x_1) + \nabla r(x_1)\right)^T (x_2 - x_1) + \frac{\mu}{2} \|x_2 - x_1\|^2$

$\Rightarrow f(x) + r(x)$ is μ-strongly convex

HW1.2: prove them. Assume convexity if needed

(a) $f(\boldsymbol{x}_2) \leq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{L}{2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_2^2$

(b) $f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \frac{1}{2L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$

(c) **Co-coercivity of the gradient:**

$$(\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1))^T(\boldsymbol{x}_2 - \boldsymbol{x}_1) \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}_2) - \nabla f(\boldsymbol{x}_1)\|_2^2$$

a) The difference between $f(x_2)$ and $f(x_1)$ can be expressed as an integral

$$f(x_2) - f(x_1) = \int_0^1 \nabla f(x_1 + t(x_2 - x_1))^T (x_2 - x_1)\, dt$$

$$= \int_0^1 \left[\nabla f(x_1 + t(x_2 - x_1)) + \nabla f(x_1) - \nabla f(x_1)\right]^T (x_2 - x_1)\, dt$$

$$f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) = \int_0^1 \left[\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\right]^T (x_2 - x_1)\, dt$$

by taking absolute value on both side

$$0 \leq \left| f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \right|$$

$$= \left| \int_0^1 \left[\nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1)\right]^T (x_2 - x_1)\, dt \right|$$

by $\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx$ and $|ab| \leq |a||b|$ we have

$$0 \le \left| f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \right|$$

$$\le \int_0^1 \| \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1) \| \cdot \| x_2 - x_1 \| \, dt$$

- f is L-smooth $\iff \| \nabla f(x_1 + t(x_2 - x_1)) - \nabla f(x_1) \| \le \frac{L}{2} \| t(x_2 - x_1) \|$

$$0 \le \left| f(x_2) - f(x_1) - \nabla f^T(x_1)(x_2 - x_1) \right| \le \int_0^1 \frac{L}{2} |t| \, \| x_2 - x_1 \|^2 \, dt$$

$$= \frac{L}{2} \| x_2 - x_1 \|^2$$

- f is convex $\iff$ the Bregman divergence $f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1)$ is non-negative

So absolute sign can be removed

$$0 \le f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \le \frac{L}{2} \| x_2 - x_1 \|_2^2$$

$\square$

b)

$$f(x_2) - f(x_1) = f(x_3) - f(x_1) - \left[f(x_3) - f(x_2) \right]$$

$$\overset{\text{I}}{\leq} \nabla f(x_1)^T (x_3 - x_1) - \left[\nabla f(x_2)^T (x_3 - x_2) + \frac{L}{2} \|x_3 - x_2\|_2^2 \right]$$

where I comes from convexity of f $\left[f(x_3) \geq f(x_1) + \nabla f(x_1)^T (x_3 - x_1) \right]$ and quadratic upper bound inequality that we proved in a

$$= \nabla f(x_1)^T (x_3 - x_1) - \nabla f(x_2)^T (x_3 - x_2) - \frac{L}{2} \|x_3 - x_2\|_2^2$$

$$= \nabla f(x_1)^T (x_2 - x_1 + x_3 - x_2) - \nabla f(x_2)^T (x_3 - x_2) - \frac{L}{2} \|x_3 - x_2\|_2$$

$$= \nabla f(x_1)^T (x_2 - x_1) + (\nabla f(x_1) - \nabla f(x_2))^T (x_3 - x_2) - \frac{L}{2} \|x_3 - x_2\|_2^2$$

rearrange the inequality:

$$(\nabla f(x_1) - \nabla f(x_2))^T (x_3 - x_2) - \frac{L}{2} \|x_3 - x_2\|_2^2 \overset{\text{I}}{\leq} f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1)$$

let $x_3 = x_2 - \frac{1}{L} (\nabla f(x_2) - \nabla f(x_1))$

we have: $x_3 - x_2 = -\frac{1}{L}\left(\nabla f(x_2) - \nabla f(x_1)\right)$ $\star$

$$\left(\nabla f(x_1) - \nabla f(x_2)\right)^T (x_3 - x_2) = \frac{1}{L}\left\|\nabla f(x_2) - \nabla f(x_1)\right\|_2^2 \quad \text{I}$$

from $\star$, we can write

$$-\frac{L}{2}\|x_3 - x_2\|_2^2 = -\frac{1}{2L}\left\|\nabla f(x_2) - \nabla f(x_1)\right\|_2^2 \quad \text{II}$$

$$\xRightarrow{\text{I},\,\text{II}} \frac{1}{2L}\left\|\nabla f(x_2) - \nabla f(x_1)\right\|_2^2 \le f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1) \quad \text{III}$$

$\square$

---

c) based on results in b, we can write

$$\frac{1}{2L}\left\|\nabla f(x_1) - \nabla f(x_2)\right\|_2^2 \le f(x_1) - f(x_2) - \nabla f(x_2)^T (x_1 - x_2)$$

$$\frac{1}{2L}\left\|\nabla f(x_2) - \nabla f(x_1)\right\|_2^2 \le f(x_2) - f(x_1) - \nabla f(x_1)^T (x_2 - x_1)$$

Sum them up:

$$\frac{1}{L}\left\|\nabla f(x_2) - \nabla f(x_1)\right\|_2^2 \le \left(\nabla f(x_2) - \nabla f(x_1)\right)^T (x_2 - x_1)$$

$\square$

HW1.3: define convergence rates

We want to find convergence rates of GD with constant step size for two scenarios:

1. convex and L-smooth. If we run gradient descent for k iterations with a fixed step size $\alpha \leq \frac{1}{L}$, we would have

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2k\alpha}$$

where $f^* = f(x^*)$ is the optimal value.

proof:

due to smoothness, we can write

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$$

Now let's plug in the gradient descent update by letting $y = x^+ = x - \alpha \nabla f(x)$. We then have

$$f(x^+) \le f(x) + \nabla f(x)^T (x - \alpha \nabla f(x) - x) + \frac{L}{2} \| x - \alpha \nabla f(x) - x \|_2^2$$

$$= f(x) - \alpha \| \nabla f(x) \|_2^2 + \frac{L}{2} \alpha^2 \| \nabla f(x) \|_2^2$$

$$= f(x) - \left(1 - \frac{L\alpha}{2}\right) \alpha \| \nabla f(x) \|_2^2$$

using $\alpha \le \frac{1}{L} \Rightarrow -\left(1 - \frac{L\alpha}{2}\right) = \frac{1}{2} L\alpha - 1 \le \frac{1}{2} L \left(\frac{1}{L}\right) - 1 = -\frac{1}{2}$

$$\Rightarrow \quad f(x^+) \le f(x) - \frac{1}{2} \alpha \| \nabla f(x) \|_2^2 \quad \text{I}$$

from convexity of f, we can write

$$f(x) \le f(x^*) + \nabla f(x)^T (x - x^*) \quad \text{II}$$

according to I and II, we have

$$f(x^+) \le f(x^*) + \nabla f(x)^T (x - x^*) - \frac{1}{2} \alpha \| \nabla f(x) \|_2^2$$

$\Rightarrow$

$$f(x^+) - f(x^*) \le \frac{1}{2\alpha} \left(2\alpha \nabla f(x)^T (x - x^*) - \alpha^2 \| \nabla f(x) \|_2^2 - \| x - x^* \|_2^2 + \| x - x^* \|_2^2 \right)$$

$$= \frac{1}{2\alpha} \left(\| x - x^* \|_2^2 - \| x - \alpha \nabla f(x) - x^* \|_2^2 \right)$$

where $x^+ = x - \alpha \nabla f(x)$

$$f(x^+) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2 \right)$$

by summing over iterations, we get

$$\sum_{i=1}^{k} \left(f(x_i) - f(x^*) \right) \leq \sum_{i=1}^{k} \frac{1}{2\alpha} \left(\|x_{i-1} - x^*\|_2^2 - \|x_i - x^*\|_2^2 \right)$$

$$= \frac{1}{2\alpha} \left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2 \right)$$

$$\leq \frac{1}{2\alpha} \left(\|x_0 - x^*\|_2^2 \right)$$

f is decreasing on every iteration, the we can conclude that

$$f(x_k) - f(x^*) \leq \frac{1}{k} \sum_{i=1}^{k} \left(f(x_i) - f(x^*) \right)$$

$$\leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k} \qquad \square$$

$\boxed{\text{2- } \mu\text{-Strongly convex and } L\text{-Smooth.}}$

with $\alpha = \frac{2}{\mu + L}$, we have $\|x_k - x^*\|_2^2 \leq \left(1 - \frac{2}{1 + \frac{L}{\mu}}\right)^{2k} \|x_0 - x^*\|_2^2$

proof: there is a theorem which implies "If f is L-smooth and μ-strongly convex, GD algorithm $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ with step sizes $\alpha_k \leq \frac{2}{\mu+L}$ satisfies

$$f(x_k) - f^* \leq \frac{L}{2} \prod_{i=0}^{k-1} \left(1 - 2\alpha_i \frac{\mu L}{\mu+L}\right) \|x_0 - x^*\|_2^2$$

where its proof is available in the literature. Now by considering this theorem with $\alpha_k = \alpha$ for all k:

$$f(x_k) - f^* \leq \frac{L}{2} \underbrace{\left[1 - 2\alpha \frac{\mu L}{\mu+L}\right]^k}_{} \|x_0 - x^*\|_2^2$$

let $\alpha = \frac{2}{\mu+L} \Rightarrow$

$$1 - 2\alpha \frac{\mu L}{\mu+L} = 1 - \frac{4\mu L}{(\mu+L)^2} = \frac{(\mu+L)^2 - 4\mu L}{(\mu+L)^2} = \frac{\mu^2 - 2\mu L + L^2}{(\mu+L)^2}$$

$$= \left[\frac{\mu - L + L - L}{\mu+L}\right]^2 = \underbrace{\left[1 - \frac{2}{1+\frac{L}{\mu}}\right]^2}_{}$$

$\square$

Problem 1.3

Define, discuss the benefits, and give examples for the different convergence rates of a sequence of updates $\{\mathbf{x}_k\}$:

(a) Sublinear

(b) Linear

(c) Superlinear

(d) Quadratic

by considering the minimization problem, the convergence to the minimizer x^* is defined as

$$\|x_k - x^*\| \leq \varepsilon$$

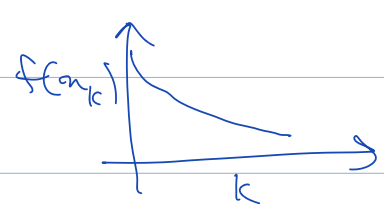
to make sure x is ε-close to the minimizer x^*

a) Sublinear convergence means k is order $O\left(\frac{1}{\varepsilon^p}\right)$

Then the equation would be

$$\|x_k - x^*\| \leq \frac{1}{k^{\frac{1}{p}}} \underbrace{\|x_0 - x^*\|}_{\text{let } R_0}$$

where p is a constant, implies sub-linear convergence



$$\frac{R_0}{k^{\frac{1}{p}}} \leq \varepsilon \Rightarrow k \geq \frac{R_0}{\varepsilon^p}$$

b) Linear convergence means k is order $\mathcal{O}\left(\log \frac{1}{\varepsilon}\right)$

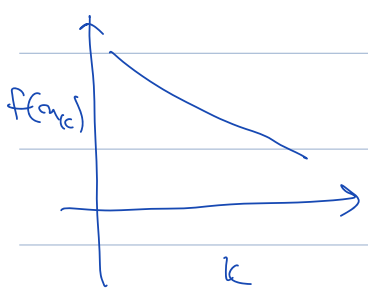
The equation

$$\|x_k - x^*\| \leq q \|x_{k-1} - x^*\|$$
$$\leq q^k \|x_0 - x^*\|$$

where $q < 1$, implies linear convergence.

$$q^k \leq \frac{\varepsilon}{R_0} \Rightarrow k \log q \leq \log \frac{\varepsilon}{R_0} \Rightarrow k \geq \frac{\log \varepsilon}{\log q} - \frac{\log R_0}{\log q}$$

$$\Rightarrow k \geq \frac{1}{|\log q|} \log \frac{1}{\varepsilon} + \frac{\log R_0}{|\log q|} = C_1 \log \frac{1}{\varepsilon} + C_2$$



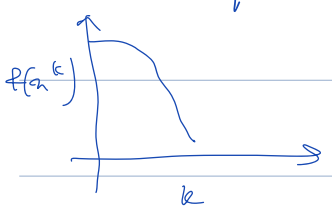
- The simple sequence $x_k = 1 + \left(\frac{1}{2}\right)^k$ converges linearly to $x^* = 1$ because

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \frac{\left(\frac{1}{2}\right)^{k+1}}{\left(\frac{1}{2}\right)^k} = \frac{1}{2}$$

c) Superlinear convergence means

$$\lim_{k \to \infty} \frac{\| x_{k+1} - x^* \|}{\| x_k - x^* \|} = 0$$

- The sequence $x_k = 1 + \left(\frac{1}{k}\right)^k$ converges superlinearly to 1



---

d) Quadratic convergence means k is order $O\left(\log\log\frac{1}{\varepsilon}\right)$

The equation

$$\| x_k - x^* \| \leq q \| x_{k-1} - x^* \|^2$$
$$\leq q^{2^k - 1} \| x_0 - x^* \|^{2^k}$$

implies quadratic convergence.

$$\Rightarrow q^{2^k - 1} R_0^{2k} \leq \varepsilon \xrightarrow[\text{Simplification}]{\text{Some}} k \geq C_1 \log\log\frac{1}{\varepsilon} + C_2$$

- The sequence $x_k = 1 + \left(\frac{1}{k}\right)^{2^k}$ converges quadratically to 1.

$$\frac{\| x_{k+1} - x^* \|}{\| x_k - x^* \|^2} = \frac{\left(\frac{1}{k+1}\right)^{2^{k+1}}}{\left(\frac{1}{k}\right)^{(2^k)^2}} = \left(\frac{k}{k+1}\right)^{2^{k+1}} \leq 1$$

HW1.4: Consider

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N}\sum_{i\in[N]} f_i(x_i) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

for $\boldsymbol{A} \in \mathbb{R}^{p\times N}$ and $\boldsymbol{x} = [x_1, \ldots, x_N]^T$.

(a) Assume strong-convexity and smoothness on f. How would you solve this problem when $N = 1000$?

(b) What if $N = 10^9$?

(c) Can we use Newton's method for $N = 10^9$? Try efficient method for computing $\nabla^2 f(\boldsymbol{x}_k)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \le p \ll N$.

(d) Now, add twice differentiable $r(\boldsymbol{x})$ to the objective and solve (a)-(c).

* duality - convergence rate GD and newton - discuss cost and rate - d:

(a) We can solve with gradient descent method.

(b) We can solve by stochastic gradient descent with mini batches so that iterations are easier but more it. will be needed

(c) i) No because we need to calculate preconditioning matrix depending on the Hessian like $P = \begin{bmatrix} \nabla^2 f & A \\ A & \nabla^2 f \end{bmatrix}$ where the size is $N+p \times N+p$, in this case we need to find inverse of it with dimensions very high, so computationally very hard.

ii) Perhaps we can replace x with a dual variable v of size p, so that the inverse matrix calc. for the Hessian will get easier.

f) We can add $r(x) = \lambda \|x\|_2^2$, and try to change the condition number. In this way, we can get faster convergence.

HW1.5: prove $(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x}-\boldsymbol{y}) \geq \frac{\mu L}{\mu + L}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \frac{1}{\mu+L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$

Let's define $\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$. Now, we should prove that ① $\phi(x)$ is convex and ② $\phi(x)$ is $(L-\mu)$-smooth and thus it has coerive gradient.

proof ①: as f is μ-strongly convex, hence $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex $\square$

proof ②: show $\phi(y) - \phi(x) - \nabla\phi(x)^T(y-x) - \frac{L-\mu}{2}\|y-x\|_2^2 \leq 0$

$$\Rightarrow \phi(y) - \phi(x) - \nabla\phi(x)^T(y-x) - \frac{L-\mu}{2}\|y-x\|_2^2$$

$$= f(y) - \frac{\mu}{2}\|y\|_2^2 - f(x) + \frac{\mu}{2}\|x\|_2^2 - \nabla\phi(x)^T(y-x) - \frac{L-\mu}{2}\|y-x\|_2^2$$

$$= f(y) - f(x) + \frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) - \nabla\phi(x)^T(y-x) - \frac{L-\mu}{2}\|y-x\|_2^2$$

$$\overset{(a)}{\leq} \nabla f(x)^T(y-x) + \frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) - \nabla\phi(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$

where (a) implies from the fact that f is L-smooth so

$$f(y) - f(x) \leq \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$$

$$\phi(x) = f(x) - \frac{\mu}{2}\|x\|_2^2 \overset{\nabla}{\Longrightarrow} \nabla\phi(x) = \nabla f(x) - \mu x, \text{ thus}$$

$$\nabla f(x)^T(y-x) + \frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) - \nabla\phi(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$

$$= \nabla f(x)^T(y-x) + \frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) - (\nabla f(x) - \mu x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$

$$= \frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) + \mu x^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$

$$= \underline{\frac{\mu}{2}\left(\|x\|_2^2 - \|y\|_2^2\right)} + \mu x^T y - \underline{\mu\|x\|_2^2} + \underline{\frac{\mu}{2}}\left(\|y\|_2^2 - 2x^Ty + \|x\|_2^2\right)$$

$$= 0$$

$$\Rightarrow \phi(y) - \phi(x) - \nabla\phi(x)^T(y-x) - \frac{L-\mu}{2}\|y-x\|_2^2 \leq 0$$

$\Rightarrow$ We proved ① and ② $\Rightarrow$ by theorem of convex-smooth function, $\phi(x)$ has coerive gradient

$$(\nabla\phi(x) - \nabla\phi(y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla\phi(x) - \nabla\phi(y)\|_2^2$$

Now, we substitute $\nabla\phi(x) = \nabla f(x) - \mu x$

$$\Rightarrow (\nabla f(x) - \nabla f(y) - \mu(x-y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y) - \mu(x-y)\|_2^2$$

$\Rightarrow$

$$(\nabla f(x) - \nabla f(y))^T(x-y) - \mu\|x-y\|_2^2 \geq$$

$$\frac{\|\nabla f(x) - \nabla f(y)\|_2^2 - 2\mu(\nabla f(x) - \nabla f(y))^T(x-y) + \mu^2\|x-y\|_2^2}{L-\mu}$$

$$\left(1 + \frac{2\mu}{L-\mu}\right)(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{\|\nabla f(x) - \nabla f(y)\|_2^2}{L-\mu} + \left(\frac{\mu^2}{L-\mu} + \mu\right)\|x-y\|_2^2$$

by dividing both sides of inequality by $\left(1+\frac{2\mu}{L-\mu}\right)$, we give the result

$$(\nabla f(x)-\nabla f(y))^T(x-y) \geq \frac{\mu L}{\mu+L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x)-\nabla f(y)\|_2^2$$

$\square$